

APPARATUS AND METHOD FOR SPATIO – TEMPORAL ALIGNMENT OF IMAGE SEQUENCES

REFERENCE TO COMPUTER PROGRAM LISTING APPENDIX.

[0001] Computer program listing appendices are submitted herewith on one compact disc and one duplicate compact disc. The total number of compact discs including duplicates is two. The files on the compact disc are text files in hexadecimal format. Their names, dates of creation, directory locations, and sizes in bytes are:

[0002] “Method” directory containing file seq2seq.hex of May 3, 2001 and of length 291,460 bytes.

“Movie” directory containing file movie.hex of May 3, 2001 and of length 9,706,798 bytes.

[0003] The material on the compact discs is incorporated by reference herein.

FIELD OF THE INVENTION

[0004] The present invention relates to apparatus and methods for sequence-to-sequence alignment.

BACKGROUND OF THE INVENTION

[0005] State of the art methods for sequence-to-sequence alignment, and other relevant technologies, are described in:

- [1] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In European Conference on Computer Vision, pages 237-252, 1992.
- [2] J.R. Bergen, P.J. Burt, R. Hingorani, and S. Peleg. A three frame algorithm for estimating two-component image motion. IEEE Trans. on Pattern Analysis and Machine Intelligence, 14:886-896, September 1992.
- [3] P.J. Burt and E.H. Adelson The laplacian pyramid as a compact image code. IEEE Transactions on Communication, 31:532-540, 1983.

- [4] Y. Caspi and M. Irani. A step towards sequence-to-sequence alignment. In IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, June 2000.
- [5] Olivier Faugeras. Three-Dimensional Computer Vision - A Geometric Viewpoint. MIT Press, Cambridge, MA, 1996.
- [6] M.A. Fischler and R.C. Bolles. Ransac random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In Communications of the ACM, volume 26, 1981.
- [7] F.R. Hampel, P.J. Rousseeuw, and W.A. Stahel E. Ronchetti. Robust Statistics: The Approach Based on Influence Functions. John Wiley, New York, 1986.
- [8] M.A. Giese and T. Poggio. Synthesis and recognition of biological motion patterns on linear superposition prototypical motion sequences. In International Conference on Computer Vision, pages 73-80, 1998.
- [9] E. Grimson, P. Viola, O. Faugeras, T. Lozano-Perez, T. Poggio, and S. Teller. A forest of sensors. In International Conference on Computer Vision, pages 45-51, 1997.
- [10] K. Hanna. Direct multi-resolution estimation of ego-motion and structure from motion. In IEEE Workshop on Visual Motion, pages 156-162, Princeton, NJ, October 1991.
- [11] C.G. Harris and M. Stephens. A combined corner and edge detector. In 4th Alvey Vision Conference, 1988.
- [12] B.K.P. Horn and B.G. Schunck. Determining optical flow. Artificial Intelligence, 17:185-203, 1981.
- [13] M. Irani and P. Anandan. Parallax geometry of pairs of points for 3d scene analysis. In European Conference on Computer Vision, Cambridge, UK, April 1996.
- [14] M. Irani and P. Anandan. About direct methods. In Vision Algorithms Workshop, pages 267-277, Corfu, 1999.
- [15] M. Irani, B. Rousso, and P. Peleg. Recovery of ego-motion using region alignment. IEEE Trans. on Pattern Analysis and Machine Intelligence, 19(3):268-272, March 1997.
- [16] M. Irani, B. Rousso, and S. Peleg. Detecting and tracking multiple moving objects using temporal integration. In European Conference on Computer Vision, pages 282-287, Santa Margarita Ligure, May 1992.

- [17] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *International Journal of Computer Vision*, 12:5-16, February 1994.
- [18] R. Kumar, P. Anandan, and K. Hanna. Direct recovery of shape from multiple views: a parallax based approach. In *Proc 12th ICPR*, pages 685-688, 1994.
- [19] L. Lee R., Romano, and G. Stein. Monitoring activities from multiple video streams: Establishing a common coordinate frame, to appear in. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, (Special Issue on Video Surveillance and Monitoring), 2000.
- [20] I. Reid and A. Zisserman. Goal-directed video metrology. In *European Conference on Computer Vision*, pages 647-658, 1996.
- [21] H. Sawhney and R. Kumar. True multi-image alignment and its application to mosaicing and lens distortion correction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 450-456, 1997.
- [22] Harpreet Sawhney. 3d geometry from planar parallax. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 1994.
- [23] G. P. Stein. Tracking from multiple view points: Self-calibration of space and time. In *DARPA IU Workshop*, pages 1037-1042, 1998.
- [24] R. Szeliski and H.-Y Shum. Creating full view panoramic image mosaics and environments maps. In *Computer Graphics Proceedings, Annual Conference Series*, 8 1997.
- [25] P.H.S. Torr and A. Zisserman. Feature based methods for structure and motion estimation. In *Vision Algorithms Workshop*, pages 279-290, Corfu, 1999.
- [26] Z. Zhang, R. Deriche, O. Faugeras, and Q. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78:87-119, 1995.
- [27] M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu, *Mosaic Representations of Video Sequences and Their Applications*. *Signal Processing: Image Communication*, special issue on Image and Video Semantics: Processing, Analysis, and Application, Vol. 8, No. 4, May 1996.

[28] M. Irani and P. Anandan, *Robust Multi-Sensor Image Alignment*. IEEE International Conference on Computer Vision (ICCV), India, January 1998.

[29] Paul Viola and William M. Wells III, "*Alignment by maximization of mutual information*," International Journal of Computer Vision (IJCV), 24(2): 137-154, 1997.

[0006] The disclosures of all publications mentioned in the specification and of the publications cited therein are hereby incorporated by reference.

09862891-051001
"00T50" T6928960

[0020] Numerals in square brackets, in the following paragraphs, refer to the

[0023] It may be useful to extend image-to-image techniques into sequence-to-sequence techniques. A gradient based method is described that addresses the problem of sequence-to-sequence alignment. The particular method is a generalization of a direct image alignment method. In this particular example it may be assumed that the sequences are taken by stationary video cameras, with fixed (but unknown) internal and external parameters. Our method simultaneously estimates spatial and temporal alignment parameters without requiring prior estimation of point correspondences, frame correspondences, moving object detection, or detection of illumination variations.

[0025] An image sequence is handled as a volume of three dimensional data (x, y, t) and not as a set of two-dimensional images $\{x, y\}_{t=1}^N$. However, unlike standard 3D volumetric alignment the spatial (x, y) and the temporal (t) components have different behavior. Typically they are treated separately, and cannot be inter-mixed.

Sequence-to-Sequence vs. Image-to-Image Alignment

[0026] Sequence-to-sequence alignment can be equally applied to any other alignment methods. Hence the objective of discussion here will be the paradigm of sequence-to-sequence alignment and not just the particular method described herein. Sequence-to-sequence alignment is a generalization of image-to-image alignment. We discuss this generalization in the context of two common alignment approaches: (i) Direct methods and (ii) Feature-Based methods. Furthermore, we show that ambiguities in alignment that can result from using spatial information alone or temporal information alone can be resolved by sequence-to-sequence alignment paradigm. This is shown for both direct and feature-based sequence-to-sequence alignment.

Direct Sequence-to-Sequence Alignment

[0027] In the direct method, if there are no temporal changes in the scene, i.e., the temporal derivatives within the sequence are zero: $S_t = 0$. Therefore, for any space-time point (x, y, t) , the error term of Eq. (1) presented below reduces to:

$$\underbrace{err_{seq}(x, y, t; \vec{P})}_{seq-to-seq} = S' - S + [u, v] \begin{bmatrix} S_x \\ S_y \end{bmatrix} = I' - I + [u, v] \begin{bmatrix} I_x \\ I_y \end{bmatrix} = \underbrace{err_{img}(x, y; \vec{P})}_{img-to-img}$$

where, $I(x, y) = S(x, y, t)$ is the image frame at time t . Therefore, the SSD function of Eq. (1) presented below reduces to the image-to-image alignment objective function, averaged over all frames.

[0028] When the scene does contain temporal variations, sequence-to-sequence alignment uses more information for spatial alignment than image-to-image alignment has access to. In particular, there are ambiguous scenarios for image-to-image alignment, which sequence-to-sequence alignment can uniquely resolve. Figs. 7A-7E illustrate a case which is ambiguous for image-to-image alignment. Consider a uniform background scene with a moving ball (Fig. 7A and Fig. 7B). At any given frame all the spatial gradients are concentrated in a very small image region (the moving ball). In these cases, image-to-image alignment cannot uniquely determine the correct spatial transformation (Fig. 7C).

only the moving object information is considered (i.e., the trajectory of the airplane), then for any temporal shift, there exists a consistent spatial transformation between the two sequences, which will bring the two trajectories in Figs. 8C and 8D into alignment. Namely, in this scenario, moving object alignment (trajectory-to-trajectory alignment) may find infinitely many valid spatio-temporal transformations. Sequence-to-sequence alignment should use both static and dynamic information and will therefore uniquely resolve the ambiguity in this case, as it forces both spatial and temporal information to be brought simultaneously into alignment across the two sequences.

[0033] Pros and cons of feature-based versus direct-based methods of image-to-image alignment apply also here, e.g., it is simpler to extend the feature based approach to extract fundamental matrix across two sequences.

[0034] The above suggested approach to feature-based sequence-to-sequence alignment can only exploit dynamic changes which are due to moving objects. However, other temporal changes, such as changes in illumination, or changes that affect the features themselves (e.g., highly non rigid motion) can not be captured by this feature-based sequence-to-sequence. The direct sequence-to-sequence alignment on the other hand captures any type of dynamic changes via the temporal derivatives, without explicitly modeling these changes by features. Therefore, it can handle quite complex dynamic, as is shown in the examples of Figs. 10 - 16. In particular, dimming or brightening of the light source are often sufficient to determine the temporal alignment. Furthermore, even homogeneous image regions contribute temporal constraints in this case. This is true although their spatial derivatives are zero, since global changes in illumination produce prominent temporal derivatives.

[0035] Note that, in general, changes in illumination need not be global. For example, an outdoor scene on a partly cloudy day, or an indoor scene with spot-lights, can be exposed to local changes in illumination. Such local changes provide additional constraints on the spatial alignment parameters. An example of applying our method to sequences with only changes in illumination is shown in lines (a) - (d) of Fig. 12.

[0036] The paradigm of sequence-to-sequence alignment extends beyond any particular method. It can equally apply to feature-based matching across sequences, or other type of match measures (e.g., mutual information).

095301-05101
100T50"1602000

BRIEF DESCRIPTION OF THE DRAWINGS AND APPENDICES

[0037] The present invention will be understood and appreciated from the following detailed description, taken in conjunction with the drawings and appendices in which:

[0038] Fig. 1 is a simplified block diagram of an event, several image sequences of the event captured from differently oriented cameras and therefore each generating different and misleading perceptions of the ball trajectory, and finally, the perceived ball trajectory of a temporally refined combined sequence generated in accordance with a preferred embodiment of the present invention by spatially-temporally aligning, and combining, the various image sequences of the event;

[0039] Fig. 2A is a simplified pictorial illustration of a first input movie, from a first one of the cameras of Fig. 1, of the ball throwing event of Fig. 1;

[0040] Fig. 2B is a simplified pictorial illustration of the perceived trajectory of the ball based on the first input movie of Fig. 2A;

[0041] Fig. 3A is a simplified pictorial illustration of a second input movie, from a second one of the cameras of Fig. 1, of the ball throwing event of Fig. 1;

[0042] Fig. 3B is a simplified pictorial illustration of the perceived trajectory of the ball based on the second input movie of Fig. 3A;

[0043] Fig. 4A is a simplified pictorial illustration of a third input movie, from a third one of the cameras of Fig. 1, of the ball throwing event of Fig. 1;

[0044] Fig. 4B is a simplified pictorial illustration of the perceived trajectory of the ball based on the third input movie of Fig. 4A;

[0045] Fig. 5 is a simplified pictorial illustration of the aligned frames which, taken together, form the combined sequence of frames, with higher sampling density, shown in Fig. 1;

[0046] Fig. 6 is a diagram of a preferred method for subsampling and aligning image sequences according to a preferred embodiment of the present invention;

[0047] Figs. 7A and 7B are pictorial illustrations of corresponding frames from two unaligned sequences;

[0048] Fig. 7C is a pictorial illustrations of some possible alignments between the frames of Figs. 7A and 7B;

[0049] Figs. 7D and 7E are pictorial illustrations of two sequences of the moving ball appearing in the frames of Figs. 7A - 7B, only one of which aligns the two trajectories of the ball.

[0050] Fig. 8A is a scene of an airplane, as viewed by a first camera;

[0051] Fig. 8B is the scene of the airplane of Fig. 8A as viewed by a second camera;

[0052] Fig. 8C is the trajectory of the airplane as seen by the first camera;

[0053] Fig. 8D is the trajectory of the airplane as seen by the second camera;

[0054] Fig. 9 is a simplified flowchart illustration of a preferred image sequence alignment method constructed and operative in accordance with a preferred embodiment of the present invention;

[0055] Fig. 10 is a set of color images showing the results of applying image-to-image alignment to temporally corresponding frames, compared to sequence-to-sequence alignment of the same;

[0056] Fig. 11 illustrates sequences of color images of a moving vehicle;

[0057] Fig. 12 illustrates sequences of color images of a scene with only varying illumination;

[0058] Fig. 13 illustrates sequences of color images of a scene with non-rigid motion.

[0059] Fig. 14 illustrates a first sequence of color images of a rapidly changing scene;

[0060] Fig. 15 illustrates a second sequence of color images of the same rapidly changing scene;

[0061] Fig. 16A illustrates a superposition of corresponding frames from Figs. 14 - 15 which have not been spatio-temporally aligned; and

[0062] Fig. 16B illustrates a superposition of corresponding frames from Figs. 14 - 15 after spatio-temporal alignment.

[0063] Also attached herewith is a CD-ROM appendix which aids in the understanding and appreciation of one preferred embodiment of the invention shown and described herein.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0064] Fig. 1 is a simplified block diagram of an event, several image sequences of the event captured from differently oriented unsynchronized cameras and therefore each generating different perceptions of the ball trajectory, and finally, the perceived ball trajectory of a combined sequence with improved temporal resolution generated in accordance with a preferred embodiment of the present invention by spatially-temporally aligning, and combining, the various image sequences of the event.

[0065] Fig. 2A is a simplified pictorial illustration of a first input movie, from a first one of the cameras of Fig. 1, of the ball throwing event of Fig. 1. Fig. 2B is a simplified pictorial illustration of the perceived trajectory of the ball based on the first input movie of Fig. 2A.

[0066] Fig. 3A is a simplified pictorial illustration of a second input movie, from a second one of the cameras of Fig. 1, of the ball throwing event of Fig. 1. Fig. 3B is a simplified pictorial illustration of the perceived trajectory of the ball based on the second input movie of Fig. 3A.

[0067] Fig. 4A is a simplified pictorial illustration of a third input movie, from a third one of the cameras of Fig. 1, of the ball throwing event of Fig. 1. Fig. 4B is a simplified pictorial illustration of the perceived trajectory of the ball based on the third input movie of Fig. 4A.

[0068] Fig. 5 is a simplified pictorial illustration of the aligned frames which, taken together, form the combined sequence of frames with improved temporal resolution. The markers 200, 210 and 220 are virtual and correspond to ball positions 4, 9 and 10 respectively which were not captured by any of the 3 cameras shown in Fig. 1.

[0069] Fig. 6 is a diagram of a preferred method for subsampling and aligning image sequences according to a preferred embodiment of the present invention, where S'_0 is an original image sequence, S'_1 is subsampled by S'_0 as described herein, S'_2 is subsampled from S'_1 similarly, and so on.

[0070] Fig. 6 illustrates a preferred hierarchical spatio-temporal alignment framework. A volumetric pyramid is constructed for each input sequence, one for the

reference sequence (on the right side), and one for the second sequence (on the left side). The spatio-temporal alignment estimator is applied iteratively at each level. It refines the approximation based on the residual misalignment between the reference volume and warped version of the second volume (drawn as a skewed cube). The output of current level is propagated to the next level to be used as an initial estimate.

[0071] Figs. 7A-7E illustrate spatial ambiguities in image-to-image alignment Figs. 7A-7B show two corresponding frames from the two sequences. There are infinitely many valid image-to-image alignments between the two frames, some of them shown in Fig 7C. Figs. 7D-7E display two sequences of a moving ball. Only one of them aligns the two trajectories of the ball.

[0072] Figs. 8A-8D illustrate spatio-temporal ambiguity in alignment when using only temporal information. A small airplane is crossing a scene viewed by two cameras. The airplane trajectory does not suffice to uniquely determine the alignment parameters. Arbitrary time shifts can be compensated by appropriate spatial translation along the airplane motion direction. Sequence-to-sequence alignment, on the other hand, can uniquely resolves this ambiguity, as it uses both the scene dynamics (the plane at different locations), and the scene appearance (the static ground). Note that spatial information alone does not suffice in this case either.

[0073] Fig. 10 illustrates image-to-image alignment vs. sequence-to-sequence alignment Line 10(a) is a color superposition showing results of applying image-to-image alignment to temporally corresponding frames. Spatial alignment is inaccurate due to insufficient spatial information in any of these individual frames. Line 10(b) is a color superposition showing accurate alignment of the same frames obtained by sequence-to-sequence alignment.

[0074] Fig. 11 illustrates a scene with moving objects. Lines 11(a) and 11(b) display 4 representative frames (100,200,300,400) from the reference and second sequences, respectively. The spatial misalignment is easily noticeable near image boundaries, where different static objects are visible in each sequence. The temporal misalignment is noticeable by comparing the position of the gate in frames 400. In the second sequence it is already open, while still closed in the reference sequence. Line 11(c)

Fig. 12 illustrates a scene with varying illumination. Lines 12(a) and 12(b) display 4 representative frames (150,200,250,300) from the reference and second sequences, respectively. The temporal misalignment can be observed at frames 200 and 300, by small differences in illumination in the upper-left corner of the frame. Line 12(c) displays superposition of the representative frames before alignment (red and blue bands from reference sequence and green band from the second sequence). Line 12(d) displays superposition of corresponding frames after spatio-temporal alignment. The accuracy of the temporal alignment is evident from the hue in the upper left corner of frame 200, which is pink before alignment (frame 200 of line 12(c)) and white after temporal alignment (frame 200 of line 12(d)). The dark pink boundaries in line 12(d) correspond to scene regions observed only by the reference camera.

[0077] Figs. 14-16 illustrate a scene which constantly changes its appearance. Figs. 14 and 15 display 10 frames (20,30,...,110) from the reference and second sequences, respectively. It is difficult to tell the connection between the two sequences. The event in frames 90-110 in the reference sequence (14), is the same as the event in frames 20-40 in the second sequence (15). Fig. 16A displays superposition of the representative frames

before spatio-temporal alignment. Fig. 16B displays superposition of corresponding frames after spatio-temporal alignment. Due to the scale difference there is an overlap between the two sequences only in the upper right region of every frame. Fireworks in the non-overlapping regions appear dark pink, as they were observed only by one camera. Fireworks in the overlapping regions appear white, as they should. The recovered temporal misalignment was approximately 66 frames.

[0078] A particular objective of a preferred embodiment of the present invention is to overcome the limitations of an individual imaging process by suitably combining information generated by a plurality of imaging processes which preferably differ in their imaging qualities so as to complement one another upon combination, thereby reducing the effect of the limitations of each individual imaging process.

[0079] The differences between the imaging processes which allow the imaging processes to complement one another if the imaging processes are suitably combined may include but are not limited to one of or a combination of the following: time offset, temporal and spatial resolution, spatial position and angular orientation, zoom, and physical characteristics of the image capturing devices such as aperture dimensions, focal length, exposure time and spectral response.

[0080] For example, as shown in Fig. 1, a plurality of sequences of images are received, such as three sequences 50, 90 and 120 in the illustrated embodiment, captured by different image capturing devices typically imaging the same scene 180. In the illustrated embodiment, the image capturing devices are cameras I, II and III also designated by reference numerals 20, 30 and 40 respectively. Each sequence, as shown, comprise an ordered multiplicity of images. For example, the sequence imaged by camera 20 is shown, for simplicity, to include three images 60, 70 and 80.

[0081] A particular advantage of a preferred embodiment of the invention as shown and described herein is that, as illustrated in Fig. 1, individual imaging processes, each of which have limitations, can due to those limitations represent an event so imperfectly as to be genuinely misleading. For example, as shown in Fig. 1, due to the insufficient temporal sampling employed by each of the three imaging devices 20, 30 and 40, none of the devices succeeds in correctly representing the S-shaped trajectory actually followed

by the ball in the true scene 180. The first camera 20, as shown in Figs. 2A - 2B, perceives a straight-line trajectory because it images the ball only at positions 1, 6 and 11 which happen to fall roughly along a straight line. The second camera 30, as shown in Figs. 3A - 3B, perceives a banana-shaped trajectory because it images the balls only at positions 2, 5 and 8 which happen to fall roughly along a banana-shaped curve. The third camera 40, as shown in Figs. 4A - 4B, perceives a inverted banana-shaped trajectory because it images the balls only at positions 3, 7 and 11 which happen to fall roughly along an inverted banana-shaped curve.

[0082] It is appreciated that the embodiment of Fig. 1 is only one example of the applicability of the present invention. More generally, the effects of a very wide variety of imaging process limitations may be reduced by the image sequence combination methods shown and described herein. Fig. 1 pertains to compensating for the specific limitation of temporal resolution however this limitation is illustrated in Fig. 1 only by way of example.

[0083] To overcome the limitations of individual imaging processes by suitably combining information generated by a plurality of differently limited imaging processes, the image sequences generated by the various individual imaging processes are typically brought into spatial and temporal alignment as shown in Fig. 5. The three image sequences of Fig. 1 are misaligned both spatially and temporally although more generally, image sequences can also be misaligned either only spatially or only temporally.

[0084] In Fig. 5, the sequences are pictorially shown to be spatially aligned as evidenced by the three different orientations of the frames of type I, originating from camera I in Fig. 1, the frames of type II, originating from camera II, and the frames of type III, originating from camera III. As shown, the frames of type I are skewed such that their upper right hand corners has been pivoted upward, the frames of type III are skewed such that their upper left hand corners has been pivoted downward, and the frames of type II are skewed intermediately between the frames of types I and III. The particular spatial misalignment illustrated pictorially in Fig. 1 is merely illustrative and is not intended to be limiting. Computational methods for effecting the alignment shown

pictorially in Fig. 5 are described in detail herein.

[0085] In Fig. 5, the sequences are pictorially shown to be temporally aligned as evidenced by the ordering of frames within the combined sequence of frames and by the three imaginary temporal spacers 200, 210 and 220 interspersed therewithin which pictorially represent the conservation of correct intervals in the combined sequence. In Fig. 1, ball positions 1 - 11 respectively are imaged by the following cameras: I, II, III, none, II, I, III, II, none, none, III. Therefore, the frames corresponding to these ball positions appear within the combined sequence of Fig. 5 in exactly this order. The nonuniform intervals between the frames in the combined sequence, depicted by means of imaginary temporal spaces 200, 210 and 220 in Fig. 5, occur because of missing ball positions at those temporal locations, i.e. ball positions 4, 9 and 10 which are not captured by any of the imaging processes of Fig. 1.

[0086] It is appreciated that the ball positions 1 - 11 in Figs. 1 - 5, which are equally spaced along the temporal dimension, are illustrated only for ease of explanation, in order to pictorially represent the temporal resolutions of the component imaging processes and how these are merged. The choice of these ball positions and not others for this illustratory purpose has no significance.

[0087] According to a preferred embodiment of the present invention, each image sequence is handled as a representation or sample of a three dimensional volume of data (x,y,t) corresponding to the event as a whole, rather than merely as a collection of two-dimensional (x,y) frames. The three dimensional volume of output data (x, y, t), generated by combining several image processes and representing, typically better than any of the image processes in isolation, the entirety of an event in space and in time, is termed herein the "space-time volume" or "spatio-temporal volume" of the event. Each image sequence corresponds to a space-time volume ("the space-time volume corresponding to the image sequence") which is an incomplete and sometimes inaccurate representation of the true space-time volume representing the entirety of the event. More generally, the dimension of the space-time volume representing an event imaged by one or more n-dimensional imaging processes is n+1. To give a simple example, a circular disc which, in a given event, is moving in space along a straight line perpendicular to the

disc's surface generates a cylinder in the three-dimensional (x,y,t) space-time volume representing that event. Another example is that a line which, in a given event, is translating through space generates a plane in the three-dimensional (x,y,t) space volume representing that event.

[0088] The alignment process shown and described herein preferably spatially and/or temporally aligns the spatio-temporal volumes corresponding to each of the several imaging processes capturing the event in question as shown by way of example in Fig. 5. The output of the alignment process shown and described herein is an output indication relating at least one point in a first space-time volume corresponding to a first sequence (more typically a multiplicity of points which may correspond to each and every pixel within each of a plurality of frames in that first space-time volume) to a corresponding number of points in at least a second space-time volume corresponding to a second sequence. Typically a dense correspondence between the space-time volumes of the various image sequences is generated including providing the above output for each of the sequences i.e. each sequence serves in turn as the first sequence. Preferably, subpixel, sub-frame-rate resolution alignment between image processes is obtained.

[0089] According to a preferred embodiment of the present invention, spatial-temporal misalignment between image processes capturing substantially the same event is derived from the outputs of the image processes themselves, e.g. the frames captured by the various image processes, as described in detail below. Therefore, it is typically unnecessary to receive external inputs defining the spatial misalignment or the temporal misalignment between image processes and therefore the image capturing devices need not be calibrated either individually or to one another in the spatial dimension and need not be synchronized individually or to one another in the temporal dimension. To derive misalignment between a plurality of image processes, a global alignment in time and/or space between representative portions of the space-time volumes corresponding to at least a pair of the plurality of image processes to be aligned is found by a matching or correlating process described herein in which spatial-temporal variations, or at least temporal variations, occurring within at least one of the space-time volumes are matched to spatial-temporal variations, or at least temporal variations,

[0090] The Sequence Brightness Error is now described with reference to the core, inner-loop step in the method of Fig. 6.

[0092] In the current implementations \bar{P} was assumed to be a parametric transformation. However, the presented framework is more general, and is not restricted to parametric transformations alone.

[0094] In the prior art, the spatial alignment parameters were recovered directly from image brightness variations, and the coarse-to-fine estimation was done using a Gaussian image pyramid. This is generalized to recover the spatial and temporal alignment parameters directly from sequence brightness variations, and the coarse-to-fine estimation is done within a *volumetric sequence pyramid*. The multi-scale analysis is done both in *space* and in *time* as shown pictorially in Fig. 6.

[0095] A global constraint on \vec{P} is obtained by minimizing the following SSD objective function:

where: $u = u(x, y, t; \vec{P}), v = v(x, y, t; \vec{P}), w = w(x, y, t; \vec{P})$ The parameter vector \vec{P} is estimated e.g. using the Gauss-Newton minimization technique. To get a term which is

explicit in the unknown parameters, linearize the term in Eq. 1 with respect to the parameter vector P to obtain:

$$ERP(\vec{P}) = \sum_{x,y,t} [S'(x,y,t) - S(x,y,t) + \nabla S(x,y,t) J_p \vec{P}]^2 \quad (2)$$

where $\nabla S = [S_x S_y S_t] = \left[\frac{\partial S}{\partial x} \frac{\partial S}{\partial y} \frac{\partial S}{\partial t} \right]$ denotes a spatio-temporal gradient of the sequence S ,

and J_p denotes the Jacobian matrix $J_p = \begin{bmatrix} \frac{\partial x'}{\partial P_1} & \frac{\partial x'}{\partial P_n} \\ \frac{\partial y'}{\partial P_1} & \frac{\partial y'}{\partial P_n} \\ \frac{\partial t'}{\partial P_1} & \frac{\partial t'}{\partial P_n} \end{bmatrix}$.

P is estimated by least-squares minimization. We solve the following “normal equations”:

$$\sum_{x,y,t} J_p \nabla S(x,y,t) (J_p \nabla S(x,y,t))^T \vec{P} = \sum_{x,y,t} [S'(x,y,t) - S(x,y,t)] J_p \nabla S(x,y,t) \quad (3)$$

or in short notations:

$$\sum_{x,y,t} J_p \nabla S (J_p \nabla S)^T \vec{P} = \sum_{x,y,t} [S' - S] J_p \nabla S \quad (4)$$

A different linearization (with respect to (x,y,t)) is possible as well:

$$e(x,y,t; \vec{P}) = S'(x,y,t) - S(x,y,t) + [uvw] \nabla S(x,y,t) \quad (3)$$

and $\nabla S = [S_x S_y S_t] = \left[\frac{\partial S}{\partial x} \frac{\partial S}{\partial y} \frac{\partial S}{\partial t} \right]$ denotes a spatio-temporal gradient of the sequence S . Eq.

(3) directly relates the unknown displacements (u, v, w) to measurable brightness variations within the sequence. To allow for large spatio-temporal displacements (u, v, w) , the minimization of Equations (1) or (3) is done within an iterative-warp coarse-to-fine framework as described herein.

[0096] In $\vec{P} = (\vec{P}_{spatial}, \vec{P}_{temporal})$ may be chosen to be a parametric transformation. Let

$\vec{p} = (x, y, 1)^T$ denote the homogeneous spatial coordinates of a spatio-temporal “voxel” (x, y, t) . Let H be the 3 x 3 matrix of the spatial parametric transformation between the two

$$u(x, y, t) = \frac{H_1 \bar{p}}{H_3 \bar{p}} - x, \text{ and } v(x, y, t) = \frac{H_2 \bar{p}}{H_3 \bar{p}} - y.$$

$\vec{P} = [h_{11} \ h_{12} \ h_{13} \ h_{21} \ h_{22} \ h_{23} \ d_1 \ d_2]$, i.e., eight unknowns. The individual voxel error of Eq. (3) becomes:

which is linear in all unknown parameters.

Each spatio-temporal “voxel” (x, y, t) provides one constraint:

The 2D projective transformation is not linear in the unknown parameters, and hence preferably undergoes some additional manipulation. To overcome this non-linearity, Eq. (4) is multiplied by the denominator $(H_3 \bar{p})$, and renormalized with its current estimate from

$$e_{new}(x, y, t; \vec{P}) = \frac{H_3 \vec{p}}{\hat{H}_3 \vec{p}} \cdot e_{old}(x, y, t; \vec{P}), \quad (5)$$

[0099] Let \hat{H} and \hat{d} be the current estimates of H and d , respectively. Substituting $H = \hat{H} + \delta H$ and $d = \hat{d} + \delta d$ into Eq. (5), and neglecting high-order terms, leads to a new error term, which is linear in all unknown parameters (δH and δd). In addition to second order terms (e.g, $\delta H \delta d$), the first order term $\hat{d} \delta H_3$ is also negligible and can be ignored.

Spatio-Temporal Volumetric Pyramid (Optional)

[0102] The Gaussian image pyramid [3] is generalized to a Gaussian sequence (volumetric) pyramid. Alternatively, a Laplacian pyramid can be used. The highest resolution level is defined as the input sequence. Consecutive lower resolution levels are obtained by low-pass filtering (LPF) both in space and time, followed by sub-sampling by a factor of 2 in all three dimensions x , y and t . Thus, for example, if one resolution level of the volumetric sequence pyramid contains a sequence of 64 frames of size 256 x 256 pixels, then the next resolution level contains a sequence of 32 frames of size 128 x 128, etc. [4] Preferably the minimal size of the spatio – temporal volume, at the highest level, is no less

than 100 pixels x 100 pixels x 10 frames for projective transformations, and 50 pixels x 50 pixels x 5 frames for simple transformations such as translations.

Summary of the process

[0103] A preferred iterative-warp coarse-to-fine estimation process is schematically described in Fig. 9, and is summarized herein:

1. (Optional) Construct two spatio-temporal volumetric pyramids S_0, S_1, \dots, S_L and S'_0, S'_1, \dots, S'_L , one for each input sequence S and S' , where $S_0 := S$ and $S'_0 := S'$. Set $\bar{P} := \bar{P}_0$ (usually the identity transformation).

2. For every resolution level, $l = L..0$, (or for only a single resolution level $l = 0$) do:

(a) Warp S'_l using the current parameter estimate: $\hat{S}'_l := \text{warp}(S'_l; \bar{P})$.

(b) Refine \bar{P} according to the residual misalignment between the reference S_l and the warped \hat{S}'_l .

(c) Repeat steps (a) and (b) until $\|\Delta P\| < \varepsilon$.

3. (Optional) Propagate \bar{P} to the next pyramid level $l - 1$, and repeat the steps (a)-(c) for S_{l-1} and S'_{l-1} .

[0104] The resulting \bar{P} is the spatio-temporal transformation, and the resulting alignment is at sub-pixel spatial accuracy, and sub-frame temporal accuracy. Results of applying this algorithm to real image sequences are shown below.

[0105] In our experiments, two different interlaced CCD cameras (mounted on tripods) were used for sequence acquisition. No synchronization what so ever was used. Typical sequence length is several hundreds of frames. Lines (a)-(d) in Fig. 11 shows a scene with a car driving in a parking lot. The two input sequences line 11(a) and line 11(b) were taken from two different windows of a tall building. Line 11(c) displays superposition of representative frames, generated by mixing the red and blue bands from the reference sequence with the green band from the second sequence. This demonstrates the initial misalignment between the two sequences, both in time and in space. Note the different timing of the gate being lifted (temporal misalignment), and misalignment of static scene parts, such the parked car or the bushes (spatial misalignment). Line 11(d) shows the

superposition after applying spatio-temporal alignment. The second sequence was spatio-temporally warped towards the reference sequence according to the computed parameters. The recovered spatial affine transformation indicated a translation on the order of a 1/5 of the image size, a small rotation, a small scaling, and a small skew (due to different aspect ratios of the two cameras). The recovered temporal shift was 46.63 frames. Therefore, opposite fields at distance 46 frames were mixed together when applying the color superposition.

[0106] In Fig. 12, the sequences (a)-(d) illustrate that dynamic information cues are not restricted to independent object motion. A light source was brightened and then dimmed down, resulting in observable illumination variations in the scene. The cameras were imaging a picture on a wall from significantly different viewing angles, inducing a significant perspective distortion. Line (a) and line (b) show a few representative frames from two sequences of several hundred frames each. The effects of illumination are particularly evident in the upper left corner of the image. Note the difference in illumination in frame 200 of the two sequences- frame 200 in line 12(a) and frame 200 in line 12(b). Line 12(c) shows a superposition of the representative frames from both sequences before spatio-temporal alignment. Line 12(d) shows superposition of corresponding frames after spatio-temporal alignment. The correctness of the temporal alignment is evident from the hue in the upper left corner of frame 200, which is pink before alignment (frame 200 in line 12(c)) and white after temporal alignment (frame 200 in line 12(d)). The accuracy of recovered temporal offset (21.32 frames) was verified (up to 0.1 frame time) against the ground truth. The verification was implemented by imaging a small object (a tennis ball) that moves very fast. The objects was viewed by three fields only (not included in the part that was used to alignment). The tennis ball location enables to verify manually that correct field -to field temporal corresponds. Furthermore, the phase differences of these locations (3 in each sequences) produce sub field accuracy "ground truth". We manually distinguish between 5 phase of $\left\{0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}\right\}$ of field time. Therefore we can verify our results up to 0.1 frame time. This suggests that the temporal offset was a component of 0.6 of field time.

[0107] In Fig. 13 the sequences (a)-(d) illustrate a case where the dynamic changes within the sequence are due to non-rigid motion (a flag blowing in the wind). Line 13(a) and line 13(b) show two representative frames out of several hundred. Line 13(c) shows a superposition of the representative frames from both sequences before spatio-temporal alignment. Line 13(d) shows superposition of corresponding frames after spatio-temporal alignment. The recovered temporal offset was 31.43 frames. Image-to-image alignment performs poorly in this case, even when applied to temporally corresponding frames, as there is not enough spatial information in many of the individual frames. This is shown in Fig. 10. We applied image-to-image alignment to all temporally corresponding pairs of fields, (odd fields from one camera with even fields from the second camera as the computed time shift (31.4) is closer to 31.5 than the 31.0) Only 55% of corresponding frames converged to accurate spatial alignment. The other 45% suffered from noticeable spatial misalignment. A few representative frames (out of the 45% of misaligned pairs) are shown in Fig. 10, line (a). These pairs were well aligned by sequence-to-sequence alignment (Fig. 10, line (b)).

[0108] Figs. 14-16 illustrates that temporal changes may include changes in appearance of the entire scene. The sequences show explosion of fireworks. The fireworks change their appearance (size, shape, color and brightness) drastically throughout the sequence. Figs. 14 and 15 show ten representative frames from two sequences of a few hundreds frames each. Frames 20-110 are displayed from the both sequences. The event in frames 90-110 in the reference sequence (14), is the same as the event shown in frames 20-40 in the second sequence (15). Line 16(a) displays superposition of four representative frames (80-110) before applying spatio-temporal alignment. The fireworks appear green and pink, due to the superposition of the different bands from different sequences (red and blue from one sequence and green from the other). The artificial colors are due to the mixture of misaligned firework with dark background from the other sequence. Line 16(b) displays superposition of the same five representative frames after applying spatio-temporal alignment. The fireworks are now white in the overlapping image regions, as they should be, implying good spatio-temporal alignment.

[0109] The above results were mainly qualitative. To quantify the expected accuracy of the method we applied several experiments where the exact ground truth alignment was known. First we warped a sequence using a known spatio-temporal parameters, applied our method to the warped and original sequence and compared the extracted parameters with the known ones. This produced highly accurate results. Less than 0.01 frame time temporal error, and less than 0.02 pixels spatial error. The accurate results are due to the fact that the source and warped sequences are highly correlated. The only difference in corresponding "voxels" gray level is as results of the tri-linear interpolation used when of the warping the second sequence. To create test where the noise is less correlated we split a sequence into its two fields. The two field sequences are related by known temporal and spatial parameters a temporal shift of 0.5 frame time, and temporal shift of 0.5 pixel along the Y axis. Note, that in this case the data comes from the same camera, but from completely different set of pixels (odd rows in one sequence and even rows in the other sequence). We repeated the experiment several (10) times using different sequences and different spatial models (affine, projective). In all cases the temporal error was smaller than 0.02 frame time. (i.e., the recovered time shift was between 0.48 and 0.52). The recovered in the Y-translation was smaller than 0.03 pixel (i.e., the recovered Y-shift was between 0.47 and 0.53 pixel), and the overall Euclidean error all over the image was bounded by 0.1 pixels. To include error that results from using two cameras, applied this test to pairs of sequences from different cameras.

[0110] Each sequence was split into two sequences of odd and even fields. In this case the ground truth is not given the relative change is known. That is if the time shift between odd sequences from the first camera reference cameras is δ_t then the time shift between odd sequences from the first camera and even sequence from the reference camera should be $\delta_t + 0.5$, and the same holds for spatial alignment. This also was performed several times and in all cases the temporal error was bounded by 0.05 frame time and the spatial error was bounded by 0.1 pixel.

[0111] In the embodiment shown and described herein, for simplicity, the internal and external parameters of the cameras are assumed to be fixed. This is not necessarily the case. For example, if the camera rotates but does not translate, each

camera is independently aligned to a fixed coordinate system as in mosaic based applications which are described in the following publication: [27] Subsequently, the method shown and described herein which is suited to fixed parameter cameras, can be applied. Furthermore, the method shown and described herein can also be used for applications in which at least part of each image sequence can be aligned to a single coordinate system.

[0112] The feature-based alignment shown and described herein is also suitable for computing fundamental matrices.

[0113] The code stored on the CD-ROM appended herewith is particularly suited for handling of 2D parametric transformations however the method described herein is more general and is applicable not only to 2D transformations but also to 3D transformations. The transformations to which the present invention is applicable includes but is not limited to the following transformations: the hierarchy of transformations described in the above-referenced publication by Bergen et al, transformations expressed in terms of 3D parameters including epipole, homography and shape parameterization as described in [5], and transformations expressed in terms of plain + parallax parameterization as described in the above-referenced publications by Irani and Anandan, 1996; Kuman et al, 1994; Harpreet Sawhney 1994; and Szeliski and Shum, 1997.

[0114] The error function shown herein, obtained by minimizing an SSD objective function, is merely an example and is not intended to be limiting. Alternatively, any suitable conventional similarity measures known for use in image to image alignment, such as but not limited to normalized correlation measures, described in [28] or mutual information measures, described in [29] may be employed.

[0115] It is appreciated that the use of a Gaussian pyramid is not essential. Alternatively, for example, Laplacian pyramids or directional derivatives may be employed.

[0116] Sequence to sequence alignment may also be used in other applications which involve combining information. For example, several cameras may be imaging a sports event simultaneously from different angles. An occurrence, such as a goal, may be

marked manually otherwise detected in one of the sequences and, by sequence-to-sequence alignment, identified also in all of the other sequences

[0117] Alignment is also useful in applications which involve recognition of a particular event or object within a plurality of sequences. For example, in commercial advertising applications, it is sometimes desired to identify an event, such as a moving logo. The similarity measure shown and described herein may be used to facilitate this recognition task.

[0118] The CD-ROM appended herewith comprises a preferred software embodiment of the present invention. The CD-ROM has two directories: "method" and "movie". "Method" stores a hex file which can be used to create two programs whose file names are "subsamplemovie" and "matchmovies" respectively. To create the programs, do as follows:

- 1) Provide a Silicon Graphics computer model SGI O2 (R5000) with an IRIX 6.3 operating system.
- 2) Copy the file Seq2seq.hex in the \method folder stored in the appended CD-ROM into a temporary directory.
- 3) Unhex the computer listing Seq2seq.hex mentioned in step 2 using HEX IT V1.8 or greater by John Augustine, 3129 Earl St., Laureldale, Pa 19605 creating file Seq2seq.zip.
- 4) Decompress the file Seq2seq.zip using gunzip version 1.2.4 available at www.gzip.org from gzip organization, extracting all files into a temporary directory.
- 5) Compile the source code stored in the temporary directory created in step 4 using a C++ compiler version 7.1 using the file "makefile". The resulting two applications will be created:
 - a) SubsampleMovie
 - b) MatchMovies

[0119] The inputs to these programs are suitably formatted movies. Two examples of suitably formatted movies can be created as follows:

- 1) Copy the file Movie.hex in the \movie folder stored in the appended CD-ROM into a temporary directory.

- [0120] The "subsamplemovie" program is a software implementation of a preferred movie subsampling procedure. The "matchmovies" program is a software implementation of a preferred procedure for aligning movies. A preferred method for running these programs to obtain an alignment in accordance with a preferred embodiment of the present invention is as follows:

- [0121] The command may be: subsampleMovie -m movie_name, where movie_name is the name of the first input movie. The output of this process is termed herein "the first level subsample of the first input movie".

- c. Run "subsamplemovie" on the second level subsample of the first input movie. The output of this process is termed herein "the third level subsample of the first input movie". For the purposes of illustration, three levels of subsampling are described herein although the number of levels of subsampling depends on image size, application and other situational parameters and may, for example, be 4 or 5.

- 33

movie.

e. Run "matchmovies" on the third level subsamples of the first and second input movies. The command may be:

matchMovies <gm1_3> <gm2_3> <type> ,

where gm1_3 and gm2_3 are the file names identifying the third level subsamples of the first and second input movies respectively.

[0122] "Type" is the type of transformation which is desired, e.g. translation (t), affine (a), quadratic (q) or projective (p).

[0123] The output of this process is termed herein P and comprises a spatial component (transformation) and a temporal component (timeshift) of the recovered alignment between the two input sequences to the process.

f. Run "matchmovies" on the second level subsamples of the first and second input movies. The command may be:

matchMovies <gm1_2> <gm2_2> <type> <2deltat> <stfn> .

where:

gm1_2 and gm2_2 are the file names identifying the second level subsamples of the first and second input movies respectively,

type is t or a or q or p as described above,

2deltat is twice the temporal component of the output of step (e), and

stfn is the spatial transformation file name, i.e. the file name of the scaled spatial transformation component of the output of step (e). A suitable scaling comprises doubling the translation component of the spatial component of the output of step (e) and halving the projective component of the spatial component of the output of step (e).

g. Run "matchmovies" on the first level subsamples of the first and second input movies. The command may be:

matchMovies <gm1_1> <gm2_1> <type> <2deltat> <stfn> .

where:

gm1_1 and gm2_1 are the file names identifying the first level subsamples of the first and second input movies respectively,

type is t or a or q or p as described above,

2deltat is twice the temporal component of the output of step (f), and

stfn is the spatial transformation file name, i.e. the file name of the scaled spatial transformation component of the output of step (f). A suitable scaling comprises doubling the translation component of the spatial component of the output of step (f) and halving the projective component of the spatial component of the output of step (f).

h. Run "matchmovies" on the original first and second input movies.

The command may be:

matchMovies <gm1> <gm2> <type> <2deltat> <stfn>.

where:

gm1 and gm2 are the file names identifying the first and second input movies respectively,

type is t or a or q or p as described above,

2deltat is twice the temporal component of the output of step (g), and

stfn is the spatial transformation file name, i.e. the file name of the scaled spatial transformation component of the output of step (g). A suitable scaling comprises doubling the translation component of the spatial component of the output of step (g) and halving the projective component of the spatial component of the output of step (g).

The output of step (h) is the alignment of the two input movies.

For sequences (movies) which are relatively well aligned, use of the movie subsampling program may be omitted.

[0124] A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

[0125] It is appreciated that the software components of the present invention may, if desired, be implemented in ROM (read-only memory) form. The software components may, generally, be implemented in hardware, if desired, using conventional techniques.

[0126] It is appreciated that the particular embodiment described in the Appendix is intended only to provide an extremely detailed disclosure of the present invention and is not

intended to be limiting.

[0127] It is appreciated that various features of the invention which are, for clarity, described in the contexts of separate embodiments may also be provided in combination in a single embodiment. Conversely, various features of the invention which are, for brevity, described in the context of a single embodiment may also be provided separately or in any suitable subcombination.

[0128] It will be appreciated by persons skilled in the art that the present invention is not limited to what has been particularly shown and described hereinabove. Rather, the scope of the present invention is defined only by the claims that follow:

0952591-051001